Minimax lower bound: Fano's method

Tri Nguyen

Reading group - Summer 2022
Oregon State University

August 10, 2022

Let's start with an example

- Given a family of Gaussian $\mathcal{N}_d = \left\{ N(\theta, \sigma^2 I_d) | \theta \in \mathbb{R}^d \right\}$.
- God chooses a distribution $P \in \mathcal{N}_d$.
- A set of n i.i.d samples are drawn from P.
- Task: estimate the mean θ from n samples.
- Quality of estimator is measured by $\mathbb{E}\left[\left\|\theta - \widehat{\theta}\right\|^2\right]$

What could be the best performance in the worse case scenario?

- If $d = 1$, we can use Cramer-Rao lower bound.
- Sample mean estimator have the error of $\dfrac{d\sigma^2}{n}$, let's see if this error can be improved.

Setting

- From a distribution family $\mathcal{P}$, God chooses a distribution $P \in \mathcal{P}$.
- A set of n i.i.d samples X_1^n are drawn from P.
- Task: estimating $\theta(P)$ from given samples.
- Question: What would be the best performance of an ideal estimator in the worse case?
- Quality of estimator is measured by $\Phi(\rho(\theta, \widehat{\theta}))$, where:
  - $\phi := \phi(P)$ is some statistic of P
  - $\widehat{\theta} := \widehat{\theta}(X_1^n)$ is some estimator
  - $\Phi(\cdot)$ is a non-decreasing function
  - $\rho(\cdot, \cdot)$ is a semimetric

$$\mathcal{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) := \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}\left[\Phi(\rho(\theta, \widehat{\theta}))\right]$$

Finding exact $\mathcal{M}()$ is difficult, instead our attempt is to find a lower bound of it.

Sketch

$$\mathcal{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) := \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}\left[\Phi(\rho(\theta, \widehat{\theta}))\right]$$

1. Translate to probability

$$\inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}\left[\Phi(\rho(\theta, \widehat{\theta}))\right] \geq \Phi(\delta) \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{P}(\rho(\theta, \widehat{\theta}) \geq \delta)$$

2. Reduce the whole space $\mathcal{P}$ to a finite set $\{\theta_v | v \in \mathcal{V}\}$

$$\sup_{P \in \mathcal{P}} \sum_{v \in \mathcal{V}} \mathbb{P}(\rho(\theta, \widehat{\theta}) \geq \delta) \geq \frac{1}{|\mathcal{V}|} \mathbb{P}(\rho(\theta_v, \widehat{\theta}) \geq \delta)$$

3. Reduce to a hypothesis testing error (required $\mathcal{V}$ to have some properties)

$$\mathbb{P}(\rho(\theta_v, \widehat{\theta}) \geq \delta) \geq \mathbb{P}(\Psi(X_1^n) \neq v)$$

4. Finding concrete bound based on specific problems.

Theorem

Assume that there exist $\{P_v \in \mathcal{P} | v \in \mathcal{V}\}, |\mathcal{V}| \leq \infty$ *such that for* $v \neq v'$, $\rho(\theta(P_v), \theta(P_{v'})) \geq 2\delta$. *Define*

- V *to be a RV with uniform distribution over* $\mathcal{V}$, *and given* $V = v$ *we draw* $\widetilde{X}_1^n \sim P_v$.
- *For an estimator* $\widehat{\theta}$, *let* $\Psi(X_1^n) := \arg\min_{v \in \mathcal{V}} \rho(\theta(P_v), \widehat{\theta}(X_1^n))$

Then,

$$\mathcal{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi(\delta) \inf_{\Psi} \mathbb{P}(\Psi(\widetilde{X}_1^n) \neq V)$$

Some remarks:

- The X_1^n in the RHS is different from the $\widetilde{X}_1^n$ in the LHS. $\widetilde{X}_1^n$ are never observed and only served for our analysis.
- There's a trade-off in choosing δ.
- In the following, $\theta_v := \theta(P_v)$, and dependence on $\widetilde{X}_1^n$ might be omitted.

Proof

$$\mathcal{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) := \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}\left[\Phi(\rho(\theta, \widehat{\theta}))\right]$$

1. Translate to probability

$$\begin{aligned} \sup_{P \in \mathcal{P}} \mathbb{E}[\Phi(\rho(\theta, \widehat{\theta}))] &\geq \sup_{P \in \mathcal{P}} \mathbb{E}[\Phi(\delta) I(\rho(\theta, \widehat{\theta}) \geq \delta)] \\ &= \Phi(\delta) \sup_{P \in \mathcal{P}} \mathbb{P}(\rho(\theta, \widehat{\theta}) \geq \delta) \end{aligned}$$

2. Restrict to set of $\{P_v \in \mathcal{P} | v \in \mathcal{V}\}$ where $\mathcal{V}$ is some index set

$$\sup_{P \in \mathcal{P}} \mathbb{P}(\rho(\theta(P), \widehat{\theta}) \geq \delta) \geq \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathbb{P}(\rho(\theta_v, \widehat{\theta}) \geq \delta)$$

   In detail,

$$\sup_{P \in \mathcal{P}} \mathbb{P}(\rho(\theta(P), \widehat{\theta}(X_1^n)) \geq \delta) \geq \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathbb{P}(\rho(\theta(P_v), \widehat{\theta}(\widetilde{X}_1^n)) \geq \delta)$$

   where
   - X_1^n are observed data which are drawn from unknown P
   - $\widetilde{X}_1^n$ are imaginary data drawn from P_v, given that $V = v$ where $V \sim \text{Uniform}(\mathcal{V})$.

3. Now we turn to a hypothesis testing by requiring set $\{\theta_v | v \in \mathcal{V}\}$ to be a 2δ-**packing set**, i.e,

$$\rho(\theta_v, \theta_{v'}) \geq 2\delta \quad \forall v \neq v'$$

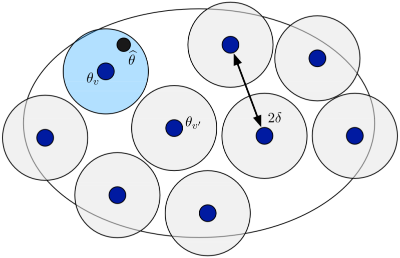

Figure: From Dr.John Duchi's notes

Recall $\Psi(X_1^n) := \arg\min_{v \in \mathcal{V}} \rho(\theta_v, \widehat{\theta}(X_1^n))$.
Since $\Psi(\widetilde{X}_1^n) \neq v \Rightarrow \rho(\theta_v, \widehat{\theta}) \geq \delta$,

$$\Rightarrow \quad \mathbb{P}(\rho(\theta_v, \widehat{\theta}) \geq \delta) \geq \mathbb{P}(\Psi(\widetilde{X}_1^n) \neq v)$$

Hence,

$$\begin{aligned}
\sup_{P\in\mathcal{P}} \mathbb{P}(\rho(\theta,\widehat{\theta}) \geq \delta) &\geq \frac{1}{|\mathcal{V}|}\sum_{v} \mathbb{P}(\rho(\theta_v,\widehat{\theta}) \geq \delta) \\
&\geq \frac{1}{|\mathcal{V}|}\sum_{v} \mathbb{P}(\Psi(\widetilde{X}_1^n) \neq v) \\
&= \mathbb{P}(\Psi(\widetilde{X}_1^n) \neq V)
\end{aligned}$$

$$\Rightarrow \inf_{\widehat{\theta}} \sup_{P\in\mathcal{P}} \mathbb{P}(\rho(\theta,\widehat{\theta}) \geq \delta) \geq \inf_{\Psi} \mathbb{P}(\Psi(\widetilde{X}_1^n) \neq V)$$

$$\Rightarrow \mathcal{M}_n(\theta(\mathcal{P}), \Phi\circ\rho) \geq \Phi(\delta)\inf_{\Psi} \mathbb{P}(\Psi(\widetilde{X}_1^n) \neq V)$$

Local Fano

Lemma (Derived from Fano inequality)

$$\inf_{\Psi} \mathbb{P}(\Psi(\widetilde{X}_1^n) \neq V) \geq 1 - \frac{I(V; \widetilde{X}_1^n) + \log 2}{\log |\mathcal{V}|}$$

Hence,

$$\mathcal{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \phi(\delta) \left(1 - \frac{I(V; \widetilde{X}_1^n) + \log 2}{\log |\mathcal{V}|} \right)$$

Mutual Information to KL

For $X_1^n \sim P_v, v \sim \mathsf{Uni}(\mathcal{V})$. Define

$$\overline{P} = \frac{1}{|\mathcal{V}|} \sum_v P_v$$

then

$$\begin{aligned}
I(V; X_1^n) = D_{\mathrm{kl}}\left(\mathbb{P}_{(V,X_1^n)} || \mathbb{P}_V \mathbb{P}_{X_1^n}\right) &= \sum_v \sum_{X_1^n} \mathbb{P}(v, x_1^n) \log \frac{\mathbb{P}(v, x_1^n)}{\mathbb{P}(v)\mathbb{P}(x_1^n)} \\
&= \sum_v \mathbb{P}(v) \sum_{X_1^n} \mathbb{P}(x_1^n | v) \log \frac{\mathbb{P}(x_1^n | v)}{\mathbb{P}(x_1^n)} \\
&= \sum_v \mathbb{P}(v) D_{\mathrm{kl}}\left(P_v || \overline{P}\right) \\
&= \frac{1}{|\mathcal{V}|} \sum_v D_{\mathrm{kl}}(P_v || \overline{P}) \\
&\leq \frac{1}{|\mathcal{V}|^2} \sum_{v,v'} D_{\mathrm{kl}}(P_v || P_{v'}) \text{(concavity of } \log\text{)}
\end{aligned}$$

How to use: A Recipe

$$\mathcal{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \phi(\delta) \left(1 - \frac{I(V; \widetilde{X}_1^n) + \log 2}{\log |\mathcal{V}|} \right) \tag{1}$$

$$I(V; \widetilde{\mathcal{X}}_1^n) \leq \frac{1}{|\mathcal{V}|^2} \sum_{v, v' \in \mathcal{V}} D_{\mathrm{kl}}(P_v || D_{v'}) \tag{2}$$

- Construct a packing set $\{\theta_v | v \in \mathcal{V}\}$ and then apply inequality (1)
  - It needs to satisfy $D_{\mathrm{kl}}(P_v || P_{v'}) \leq f(\delta)$ for some f
  - And $|\mathcal{V}|$ need to be large.
- Compute the bound $I(V; \widetilde{X}_1^n)$ as a function of δ using (2)
- Choose an optimal δ

How to use: Example

Example. Given the family $\mathcal{N}_d = \left\{N(\theta; \sigma^2 I_d) \mid \theta \in \mathbb{R}^d\right\}$. The task is to estimate the mean $\theta(P)$ for some $P \in \mathcal{N}_n$ given X_1^n samples drawn i.i.d from P. We wish to find out the lower bound of minimax error in term of mean-squared error.

Solution. Let's construct the local packing set $\{\theta_v | v \in \mathcal{V}\}$:

- Let $\mathcal{V}$ be a $1/2$-packing of unit ℓ_2-ball where $|\mathcal{V}| \geq 2^d$. It is guaranteed that such $\mathcal{V}$ exists.
- Then our $\delta/2$-packing set is $\left\{\delta v \in \mathbb{R}^d | v \in \mathcal{V}\right\}$, since

$$\|\theta_v - \theta_{v'}\|_2 = \delta \|v - v'\|_2 \geq \frac{\delta}{2} \quad \text{(since } \mathcal{V} \text{ is a 1/2-packing set)}$$

Apply our bound,

$$\begin{aligned}
\mathcal{M}_n(\theta(\mathcal{N}_d), \|\cdot\|^2) &\geq \Psi(\delta)\left(1 - \frac{I(V; X_1^n) + \log 2}{\log|\mathcal{V}|}\right) \\
&\geq \left(\frac{1}{2}\frac{\delta}{2}\right)^2 \left(1 - \frac{I(V; X_1^n) + \log 2}{\log|\mathcal{V}|}\right) \\
&= \frac{\delta^2}{16}\left(1 - \frac{I(V; X_1^n) + \log 2}{\log|\mathcal{V}|}\right)
\end{aligned}$$

And,

$$
\begin{aligned}
I(V; X_1^n) &\leq \frac{1}{|\mathcal{V}|^2} \sum_{v,v'} D_{\mathrm{kl}}(P_v^n || P_{v'}^n) \\
&= \frac{1}{|\mathcal{V}|^2} \sum_{v,v'} n D_{\mathrm{kl}} \left(N(\delta v, \sigma^2 I_d), N(\delta v', \sigma^2 I_d) \right) \\
&= n D_{\mathrm{kl}} \left(N(\delta v, \sigma^2 I_d), N(\delta v', \sigma^2 I_d) \right) \\
&= n \frac{\delta^2}{2\sigma^2} \|v - v'\|^2 \leq \frac{n\delta^2}{2\sigma^2}
\end{aligned}
$$

Let's combine these 2 inequalities above,

$$
\mathcal{M}_n(\theta(\mathcal{N}_d), \|\cdot\|^2) \geq \frac{\delta^2}{16} \left(1 - \frac{\frac{n\delta^2}{2\sigma^2} + \log 2}{d \log 2} \right)
$$

That bound's optimal value is achieved at $\delta^2 = \dfrac{(d-1)\sigma^2 \log 2}{n}$, and the optimal value is

$$
\frac{(d-1)^2 \sigma^2 \log 2}{32 dn} \Rightarrow O\left(\frac{d\sigma^2}{n} \right)
$$

Proof of the claim on packing number

Claim: There exists a $1/2$-packing set of unit ℓ_2-ball with cardinality at least 2^d.

Proof:

- A δ-packing of the set Θ with respect to ρ is a set $\{\theta_1, \ldots, \theta_M\}, \theta_i \in \Theta, i = 1, \ldots, N$ such that $\rho(\theta_v, \theta_{v'}) \geq \delta \; \forall v \neq v'$.
- Then δ-packing number is

$$M(\delta, \Theta, \rho) = \sup\{M \in \mathbb{N} : \text{there exists a } \delta\text{-packing } \{\theta_1, \ldots, \theta_M\} \text{ of } \Theta\}$$

We have

$$\begin{cases} M(\delta, \Theta, \rho) \geq N(\delta, \Theta, \rho) \\ N(\delta, \mathbb{B}, \|\cdot\|) \geq (1/\delta)^d \end{cases} \Rightarrow M(1/2, \mathbb{B}, \|\cdot\|) \geq 2^d$$

- For the first inequality, denote $\widehat{\Theta}$ be a δ-packing of Θ with size of $M(\delta, \Theta, \rho)$. Since there is no $\theta \in \Theta$ we can add to $\widehat{\Theta}$ such that $\rho(\theta, \widehat{\theta}) \geq \delta$, $\widehat{\Theta}$ is also a δ-covering of Θ.
- For the second inequality, let $\{v_1, \ldots, v_N\}$ as a δ-covering of $\mathbb{B}$, then

$$\mathsf{Vol}(\mathbb{B}(\mathbf{0}, 1)) \leq \sum_{i=1}^{N} \mathsf{Vol}(\mathbb{B}(v_i, \delta)) = N\mathsf{Vol}(\mathbb{B}(v_1, \delta)) = N\delta^d \mathsf{Vol}(\mathbb{B}(\mathbf{0}, 1))$$

Proof of the bound on mutual information

Proposition (Fano inequality)

For any Markov chain $V \to X \to \widehat{V}$, we have

$$h_2(\mathbb{P}(\widehat{V} \neq V)) + \mathbb{P}(\widehat{V} \neq V)\log(|\mathcal{V}| - 1) \geq H(V|\widehat{V})$$

where $h_2(p) = -p\log(p) - (1-p)\log(1-p)$ is entropy of a Bernoulli RV with parameter p.

Apply this proposition for V being a uniform RV over $\mathcal{V}$,

$$H(V|\widehat{V}) = H(V) - I(V;\widehat{V}) = \log|\mathcal{V}| - I(V;\widehat{V}) \geq \log|\mathcal{V}| - I(V;X)$$

Hence,

$$\begin{aligned} \log 2 + \mathbb{P}(V \neq \widehat{V})\log(|\mathcal{V}|) &> \log h_2(\mathbb{P}(V \neq \widehat{V})) + \mathbb{P}(V \neq \widehat{V})\log(|\mathcal{V}| - 1) \\ &\geq H(V|\widehat{V}) \\ &\geq \log|\mathcal{V}| - I(V;X) \end{aligned}$$

$$\Rightarrow \mathbb{P}(V \neq \widehat{V}) \geq 1 - \frac{I(V;X) + \log 2}{\log|\mathcal{V}|}$$

Proof of Fano Inequality

Let $E = 1$ be the event $V \neq \widehat{V}$, $E = 0$ otherwise. We have

$$\begin{aligned} H(V, E|\widehat{V}) &= H(V|E, \widehat{V}) + H(E|\widehat{V}) \quad \text{(chain rule)} \\ &= \mathbb{P}(E = 1)H(V|E = 1, \widehat{V}) + \mathbb{P}(E = 0)H(V|E = 0, \widehat{V}) + H(E|\widehat{V}) \\ &= \mathbb{P}(E = 1)H(V|E = 1, \widehat{V}) + H(E|\widehat{V}) \end{aligned}$$

We also have

$$\begin{aligned} H(V, E|\widehat{V}) &= H(E|V, \widehat{V}) + H(V|\widehat{V}) \\ &= H(V|\widehat{V}) \end{aligned}$$

Hence,

$$\begin{aligned} H(V|\widehat{V}) &= \mathbb{P}(E = 1)H(V|E = 1, \widehat{V}) + H(E|\widehat{V}) \\ &\leq \mathbb{P}(E = 1) \log |\mathcal{V} - 1| + H(E) \\ &= \mathbb{P}(V \neq \widehat{V}) \log(|\mathcal{V}| - 1) + h_2(\mathbb{P}(V \neq \widehat{V})) \end{aligned}$$

A variant: Distance-based Fano method

The previous derivation requires a construction of a packing set to translate to a hypothesis testing error.

$$\mathcal{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi(\delta) \inf_{\Psi} \mathbb{P}(\Psi(\widetilde{X}_1^n) \neq V)$$

The main reason is (derived) Fano's inequality:

$$\mathbb{P}(\widehat{V} \neq V) \geq 1 - \frac{I(V; X_1^n) + \log 2}{\log |\mathcal{V}|}$$

We can bound minimax without explicitly constructing packing set.

$$\mathbb{P}(\rho_{\mathcal{V}}(\widehat{V}, V) > t) \geq 1 - \frac{I(V; X_1^n) + \log 2}{\log(|\mathcal{V}| / N_t^{\max})}$$

Then,

$$\mathcal{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi\left(\frac{\delta(t)}{2}\right) \left[1 - \frac{I(X; V) + \log 2}{\log \frac{|\mathcal{V}|}{N_t^{\max}}}\right]$$

where

$$\delta(t) := \sup \{\delta | \rho(\theta_v, \theta_{v'}) \geq \delta \quad \text{for all } v, v' \in \mathcal{V} \text{ such that } \rho_{\mathcal{V}}(v, v') > t\}$$